

AD-A095 179

NAVAL RESEARCH LAB WASHINGTON DC

F/6 20/13

CROSS-ENTROPY MINIMIZATION GIVEN FULLY-DECOMPOSABLE SUBSET AND --ETC(U)

FEB 81 J E SHORE

UNCLASSIFIED NRL-MR-4430

NL

1 of 1
20 9 1 1 3

00

END

DATE

FILED

13 8 1

DTIC

621500179

January 22, 1961

DTIC

S D

THE COPY

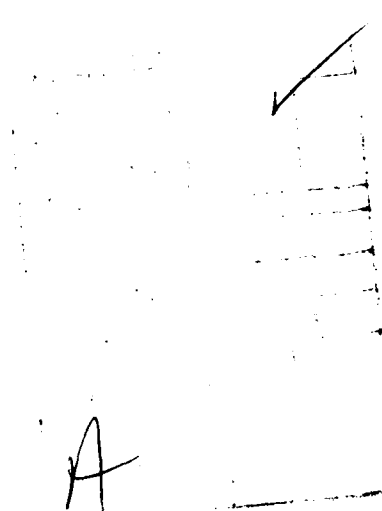
REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Memorandum Report 4430 ^v	2. GOVT ACCESSION NO. AD-A095179	3. RECIPIENT'S CATALOG NUMBER NRL - 105-450
4. TITLE (and Subtitle) CROSS-ENTROPY MINIMIZATION GIVEN FULLY-DECOMPOSABLE SUBSET AND AGGREGATE CONSTRAINTS.		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) John E. Shore		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N RR014-09-41 75-0102-0-1
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		12. REPORT DATE February 1981
		13. NUMBER OF PAGES 25
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Information theory Minimum discrimination information Cross-entropy System modeling Directed divergent Queueing networks Entropy maximization		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The principle of maximum entropy and the principle of minimum cross-entropy (minimum directed divergence, minimum discrimination information) have been applied recently to problems in queuing theory and computer system performance modeling. These information-theoretic principles estimate probability distributions based on information in the form of known expected values. In the case of queuing theory and computer system modeling, the known expected values arise from rate balance equations. This paper concerns situations in which the system state probabilities decompose into disjoint subsets and in which the known expected values are either expectations conditional on a specific subset or expectations involving aggregate subset probabilities. (Continues)		

20. Abstract (Continued)

New properties of minimum cross-entropy distributions are derived and an efficient method of computing these distributions is derived. Computational examples are included. In the case of queuing theory and computer system modeling, the disjoint subsets correspond to internal device states and the aggregate probabilities correspond to overall device states. The results of this paper apply when one has both rate balance equations for device equilibrium involving internal device state probabilities as well as rate balance equations for system equilibrium involving aggregate device state probabilities.

CONTENTS

I. INTRODUCTION	1
II. BACKGROUND	3
III. CROSS-ENTROPY MINIMIZATION WITH SUBSET AND AGGREGATE CONSTRAINTS	9
IV. COMPUTATIONAL RESULTS	18
V. ACKNOWLEDGEMENTS	20
REFERENCES	21



CROSS-ENTROPY MINIMIZATION GIVEN FULLY-DECOMPOSABLE SUBSET AND AGGREGATE CONSTRAINTS

I. INTRODUCTION

The principle of minimum cross-entropy provides a general method of inference about an unknown probability distribution $q^\dagger = q_1^\dagger, q_2^\dagger, \dots, q_N^\dagger$, when there exists a prior estimate of q^\dagger as well as new information in the form of known expected values $\sum_i f_{ki} q_i^\dagger$. The principle states that, of all the densities with the correct expected values, one should choose the posterior q with the least cross-entropy $H[q, p] = \sum_i q_i \log(q_i/p_i)$, where p is a prior estimate of q^\dagger . Cross-entropy minimization was first introduced by Kullback [1], who called it minimum directed divergence and minimum discrimination information. The principle of maximum entropy [2],[3] is equivalent to cross-entropy minimization in the special case of uniform priors. Cross-entropy minimization has a long history of applications in a variety of fields (for a list of references, see [4]). Recently, results have been obtained for spectral analysis [5], speech coding [6], pattern recognition [7], queuing theory [8],[9] and computer system modeling [10],[11]. For discussions of the the background, validity, and properties of cross-entropy minimization, see [1],[4],[12],[13].

There are a number of general algorithms for finding minimum cross-entropy distributions given arbitrary priors and arbitrary constraints [14]-[16]. Most of these algorithms are based on the Newton-Raphson method [13, Appendix A], which involves a matrix inversion during each iteration, and the computation time for all of them grows rapidly with the number of points N . Such rapid growth may be unavoidable in the case of completely general expectation functions. In this paper, we consider a less general case: one in which the known expected values are either of the forms

Manuscript submitted November 18, 1980.

$$\frac{1}{r_j^\dagger} \sum_{i \in D_j} v_{ki} q_i^\dagger \quad (1)$$

or

$$\sum_{j=1}^M u_{kj} r_j^\dagger, \quad (2)$$

where the D_j , $j = 1, 2, \dots, M$, are disjoint subsets $D_j \subset \{1, 2, \dots, N\}$, with $D_1 \cup D_2 \cup \dots \cup D_M = \{1, 2, \dots, N\}$, and where the r_j^\dagger are given by

$$r_j^\dagger = \sum_{i \in D_j} q_i^\dagger. \quad (3)$$

That is, we consider the case in which one knows expected values conditional on any of the disjoint subsets D_j as well as expected values of the distribution of aggregate probabilities r_j^\dagger . In this case, we shall show that some special properties apply and that one can solve the overall minimum cross-entropy problem by solving a minimum cross-entropy problem for each of the conditional subset distributions followed by one minimum cross-entropy problem for the aggregate distribution. If all of the subsets have equal sizes, this means that, instead of solving one N dimensional problem, one can solve M problems of dimension N/M followed by one problem of dimension M .

The results presented in this paper may be particularly useful in applications that concern queuing networks and other computer system performance models. In these applications, known expected values often arise from rate balance equations. Consider a simple example: For a $M/M/1/N$ queue with state probabilities q_k^\dagger , a typical rate-balance equation is

$$(\lambda_i + \mu_i) q_i^\dagger = \mu_{i+1} q_{i+1}^\dagger + \lambda_{i-1} q_{i-1}^\dagger,$$

where λ_i and μ_i respectively are state-dependent arrival and service rates. This equation is just a special case of the general form $\sum_i f_{ki} q_i^\dagger$. For general queuing networks, the disjoint subsets \mathcal{D}_j in (1)-(3) can correspond to internal device states and the r_j^\dagger can correspond to the probabilities of aggregate device states. The results of this paper apply when one has both rate balance equations for device equilibrium of the form (1) as well as rate balance equations for system equilibrium of the form (2).

Section II defines the notation we shall use and reviews the mathematics and justification of cross-entropy minimization. In Section III, we summarize previous results concerning minimum cross-entropy problems with subset and aggregate constraints, we prove several new properties, and we show how the minimum cross-entropy problem can be decomposed into smaller problems. Computational results comparing the full space and decomposed methods are presented in Section IV.

II. BACKGROUND

A. Notation

We use the same notation as in [4],[13]. For a more detailed discussion of technical conditions and questions related to the existence of minimum cross-entropy solutions, see [12],[13].

We use lower-case boldface Roman letters for system states, which may be multidimensional, and upper-case boldface Roman letters for sets of system states. We use lower-case Roman letters for probability densities, and upper case script letters for sets of probability densities. Thus, let \underline{x} be a state of some system that has a set \mathcal{D} of possible states. Let \mathcal{Q} be the set of all probability densities q on \mathcal{D} such that $q(\underline{x}) \geq 0$ for $\underline{x} \in \mathcal{D}$ and

$$\int_{\mathcal{D}} d\mathbf{x} \, q(\mathbf{x}) = 1 \quad . \quad (4)$$

We use a dagger \dagger to distinguish the system's unknown "true" state probability density $q^\dagger \in \mathcal{Q}$. When $\mathcal{S} \subseteq \mathcal{D}$ is some set of states, we write $q(\mathbf{x} \in \mathcal{S})$ for the set of values $q(\mathbf{x})$ with $\mathbf{x} \in \mathcal{S}$.

New information takes the form of linear equality constraints

$$\int_{\mathcal{D}} d\mathbf{x} \, q^\dagger(\mathbf{x}) f_i(\mathbf{x}) = F_i \quad , \quad (i = 1, 2, \dots, K) \quad (5)$$

for known functions f_i and known values F_i . The probability densities that satisfy such constraints always comprise a convex subset \mathcal{J} of \mathcal{Q} . We refer to the functions f_i as constraint functions and to \mathcal{J} as a constraint set. For a given constraint set there may of course be more than one set of constraint functions in terms of which it may be defined. We frequently suppress mention of a particular set of constraint functions, using the notation $I = (q^\dagger \in \mathcal{J})$ to mean that q^\dagger is a member of the constraint set \mathcal{J} and referring to I as a constraint or constraints. We use upper-case Roman letters for constraints. The results in this paper are restricted to the case of equality constraints. Results for the more general case involving inequality constraints — bounds on expected values — are discussed in [1], [4], [12], [13].

Let $p \in \mathcal{Q}$ be some prior density that is an estimate of q^\dagger obtained, by any means, prior to learning I . Priors must be strictly positive: $p(\mathbf{x} \in \mathcal{D}) > 0$ (for discussion, see [13]). Given a prior p and new information I , the posterior density $q \in \mathcal{Q}$ that results from taking I into account is chosen by minimizing the cross-entropy $H[q, p]$ in the constraint set \mathcal{J} :

$$H[q, p] = \min_{q' \in \mathcal{J}} H[q', p] \quad , \quad (6)$$

where

$$H[q, p] = \int_{\underline{D}} d\underline{x} \, q(\underline{x}) \log(q(\underline{x})/p(\underline{x})) \quad (7)$$

We introduce an "information operator" \circ that expresses (6) using the notation

$$q = p \circ I \quad (8)$$

The operator \circ takes two arguments -- a prior and new information -- and yields a posterior.

For some subset $\underline{S} \subseteq \underline{D}$ of states and $\underline{x} \in \underline{S}$, let

$$q(\underline{x} | \underline{x} \in \underline{S}) = q(\underline{x}) / \int_{\underline{S}} d\underline{x}' \, q(\underline{x}') \quad (9)$$

be the conditional density, given $\underline{x} \in \underline{S}$, corresponding to any $q \in \underline{Q}$. We use

$$q(\underline{x} | \underline{x} \in \underline{S}) = q * \underline{S} \quad (10)$$

as a shorthand notation for (9).

When \underline{D} is a discrete set of system states, densities are replaced by discrete distributions and integrals by sums in the usual way. We use lowercase boldface roman letters for discrete probability distributions, which we consider to be vectors; for example, $\underline{q} = q_1, q_2, \dots, q_N$. It will always be clear in context whether, for example, the symbol \underline{x} refers to a system state or a discrete distribution and whether r_i refers to a probability density or a component of a discrete distribution.

B. Minimum Cross-Entropy Probability Densities

Given a positive prior probability density p , if there exists a posterior that minimizes the cross-entropy (7) and satisfies the constraints (4) and (5), then it has the form

$$q(\underline{x}) = p(\underline{x}) \exp \left(-\lambda - \sum_{j=1}^K \beta_j f_j(\underline{x}) \right), \quad (11)$$

In (11), λ and β_j are Lagrangian multipliers whose values are determined by

the constraints (4) and (5). Conversely, if one can find values for λ and β_j in (11) such that the constraints (4) and (5) are satisfied, then the solution exists and is given by (11) [12]. The cross-entropy at the minimum can be expressed in terms of the Lagrangian multipliers and the expected values F_j as follows ([1, p. 38], [13]):

$$H[q, p] = -\lambda - \sum_{j=1}^K \beta_j F_j \quad (12)$$

It is necessary to choose β and the λ_j so that the constraints are satisfied. In the presence of the constraint (4), one may rewrite the remaining constraints (5) in the form

$$\int d\mathbf{x} (f_i(\mathbf{x}) - F_i) q(\mathbf{x}) = 0 \quad (13)$$

Now, if one finds values for the β_j such that

$$\int d\mathbf{x} (f_i(\mathbf{x}) - F_i) p(\mathbf{x}) \exp \left(- \sum_{j=1}^K \beta_j f_j(\mathbf{x}) \right) = 0, \quad (i = 1, \dots, M), \quad (14)$$

holds, (13) will be satisfied, and (4) can then be satisfied by setting

$$\lambda = \log \int d\mathbf{x} p(\mathbf{x}) \exp \left(- \sum_{j=1}^K \beta_j f_j(\mathbf{x}) \right) \quad (15)$$

If the integral in (15) can be performed, one can sometimes find values for the β_j from the relations

$$-\frac{\partial \lambda}{\partial \beta_j} = F_j \quad .$$

It unfortunately is usually impossible to solve this or (14) for the β_j explicitly, in order to obtain a closed-form solution expressed directly in terms of the known expected values F_j rather than in terms of the Lagrangian

multipliers. Computational methods for finding approximate solutions are, however, available ([14]-[16]).

When the prior density is uniform on D , minimizing (7) is equivalent to maximizing the entropy

$$- \int_D q(x) \log(q(x))$$

Minimum cross-entropy and maximum entropy are also equivalent when the prior is exponential in a linear combination of the constraint functions. In both cases, (11)-(15) all apply with the prior deleted.

C. Justification of Cross-Entropy Minimization

In what sense does cross-entropy minimization yield the best estimate of q^\dagger ? To answer this question, it is useful to ask what would happen if other functionals besides cross-entropy (7) were used in implementing the information operator \circ in (8). Recent work has shown that, if the operator \circ is required to satisfy certain axioms of consistent inference, and if \circ is implemented by means of functional minimization, then the principle of minimum cross-entropy follows necessarily [4]. Informally, the axioms state that different ways of taking the information I into account -- for example, in different coordinate systems -- should lead to consistent results. In terms of these axioms, the principle of cross-entropy minimization is correct in the following sense: Given a prior probability density and new information in the form of constraints on expected values, there is only one posterior density satisfying these constraints that can be chosen by functional minimization in a manner that satisfies the axioms; this unique posterior can be obtained by minimizing cross-entropy.

An additional interpretation of the sense in which $q = p \circ I$ is the best estimate of q^\dagger rests on cross-entropy's well-known [1] and unique [17]

properties as an information measure. Informally speaking, $H[q,p]$ is a measure of the "information divergence" or "information dissimilarity" between q and p . In these terms, one can interpret the principle of minimum cross-entropy as follows: Since $q = p \circ I$ minimizes $H[q,p]$, the posterior hypothesis for q^+ is as close as possible in an information-measure sense to the prior hypothesis while at the same time satisfying the new constraints I . Furthermore, in the context of cross-entropy minimization, cross-entropy satisfies the triangle equality [12],[13]

$$H[q^+, p] = H[q^+, p \circ I] + H[p \circ I, p] \quad (16)$$

Thus, the minimum cross-entropy posterior estimate of q^+ is not only logically consistent, but also closer to q^+ , in the cross-entropy sense, than is the prior p . Moreover, the difference $H[q^+, p] - H[q^+, p \circ I]$ is exactly the cross-entropy $H[p \circ I, p]$ between the posterior and the prior. Hence, $H[p \circ I, p]$ can be interpreted as the amount of information provided by I that is not inherent in p . Stated differently, $H[p \circ I, p]$ is the amount of additional distortion introduced if p is used instead of $p \circ I$. Since, for any density r there exist constraints I_r such that $r = p \circ I_r$ for any prior p , $H[r, p]$ is in general the amount of information needed to determine r when given p , or the amount of additional distortion introduced if r is used instead of p [13].

Yet another justification for using cross-entropy as a distortion measure in the context of cross-entropy minimization is provided by the "expectation matching" property [13], which states that, for an arbitrary density q^* and a density q of the general form (11), $H[q^*, q]$ is smallest when the expectations of q match those of q^* . In particular, it follows that $q = p \circ I$ is not only the density that minimizes $H[q, p]$, as already discussed, but also is the density of the form (11) that minimizes $H[q^+, q]$. Hence $p \circ I$ is not only closer to q^+ than is p -- as shown by (16) -- but it is the closest possible

density of the form (11).

III. CROSS-ENTROPY MINIMIZATION WITH SUBSET AND AGGREGATE CONSTRAINTS

In this Section, we introduce the special classes of subset and aggregate constraints and we summarize properties of $p \in I$ that apply when I consists either of subset or aggregate constraints. For the case of subset constraints only, we then show how $p \in I$ can be computed using a decomposition method that obviates solving a minimum cross entropy problem on the entire state space \mathcal{D} . Next we derive several new properties that apply when I consists of both subset and aggregate constraints and we show how the the decomposition method can be used in this case as well.

A. Subset and Aggregate Constraints

Let $\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_M$ be disjoint subsets whose union is \mathcal{D} and let $S_j = (q^+ * \mathcal{D}_j \in \mathcal{A}_j)$ be new information about the conditional density $q^+ * \mathcal{D}_j$, where $\mathcal{A}_j \subset \mathcal{D}_j$ and \mathcal{D}_j is the set of densities on \mathcal{D}_j . In particular, suppose that S_j is given by the constraints

$$\int_{\mathcal{D}_j} d\mathbf{x} [q^+ * \mathcal{D}_j](\mathbf{x}) u_{ij}(\mathbf{x}) = \bar{u}_{ij} \quad (i = 1, \dots, K_j) .$$

These can always be written in the form

$$\int_{\mathcal{D}_j} d\mathbf{x} [q^+ * \mathcal{D}_j](\mathbf{x}) s_{ij}(\mathbf{x}) = 0 \quad (i = 1, \dots, K_j) , \quad (17)$$

where $s_{ij}(\mathbf{x}) = u_{ij}(\mathbf{x}) - \bar{u}_{ij}$. Now the constraints S_j can also be written as constraints S'_j on the full density q^+ , namely $S'_j = (q^+ \in \mathcal{A}'_j)$ where $\mathcal{A}'_j \subset \mathcal{D}$. In particular, the constraints S_j in (17) correspond to the constraints S'_j given by

$$\int_{\mathcal{D}} d\mathbf{x} \, q^{\dagger}(\mathbf{x}) \, f_{ij}(\mathbf{x}) = 0 \quad (i = 1, \dots, K_j) \quad , \quad (18)$$

where

$$f_{ij}(\mathbf{x}) = \begin{cases} s_{ij}(\mathbf{x}) & , \mathbf{x} \in \mathcal{D}_j \\ 0 & , \text{otherwise} \end{cases} \quad (19)$$

We write $S = S_1 \wedge S_2 \wedge \dots \wedge S_M$ as well as $S' = S'_1 \wedge S'_2 \wedge \dots \wedge S'_M$, and we refer to S or S' as subset constraints.

When new information consists of subset constraints only, cross-entropy minimization satisfies a property known as weak subset independence [13], which states that

$$(p \circ S') * \mathcal{D}_j = (p * \mathcal{D}_j) \circ S_j \quad (20)$$

holds. Given subset constraints S and an arbitrary prior $p \in \mathcal{Q}$, there are two ways of obtaining posterior conditional densities for each subset \mathcal{D}_j : One way is to obtain a posterior $r = p \circ S'$ for the whole system and then to compute conditional posteriors $r * \mathcal{D}_j$. Another way is to obtain a conditional posterior $(p * \mathcal{D}_j) \circ S_j$ from each conditional prior using only that part of S which refers to the subset \mathcal{D}_j . Eq. (20) states that the results are the same in both cases. Furthermore, the cross-entropy $H[p \circ S', p]$ satisfies

$$H[r, p] = H[\Psi r, \Psi p] + \sum_{j=1}^M [\Psi r]_j \, H[r * \mathcal{D}_j, p * \mathcal{D}_j] \quad , \quad (21)$$

where $r = p \circ S'$, and where Ψ is a subset aggregation transformation such that, for any $q \in \mathcal{Q}$, Ψq is a discrete distribution with

$$[\Psi q]_j = \int_{\mathcal{D}_j} d\mathbf{x} \, q(\mathbf{x})$$

The transformation Ψ aggregates the states in each subset D_j . For proofs of (20)-(21), see [13]. In fact, it is easy to show by direct calculation that (21) holds for arbitrary densities $q, r \in \mathcal{Q}$.

Now consider situations in which there is information about the aggregate distribution Ψq^\dagger . In particular, let A be the constraints

$$\sum_{j=1}^M [\Psi q^\dagger]_j a_{ij} = 0 \quad (i = 1, \dots, K_a) \quad (22)$$

(Constraints with non-zero right hand sides can always be written in this form as was the case with (17).) The aggregate information A can also be expressed as information A' about the density q^\dagger . In particular, the constraints A in (22) correspond to the constraints A' given by

$$\int_D d\tilde{x} q^\dagger(\tilde{x}) a_i(\tilde{x}) = 0 \quad (i = 1, \dots, K_a) \quad (23)$$

where

$$a_i(\tilde{x} \in D_j) = a_{ij} \quad (24)$$

We refer to A' or A as aggregate constraints. Given such aggregate constraints, cross-entropy minimization satisfies a property known as subset aggregation [13], which states that

$$\Psi(p \circ A') = (\Psi p) \circ A \quad (25)$$

and

$$H[p \circ A', p] = H[\Psi(p \circ A'), \Psi p] \quad (26)$$

hold. Thus, the aggregate probabilities of the posterior $p \circ A'$ are the same as those obtained by aggregating the prior and then taking the constraints A into account. For proofs of (25)-(26), see [13].

When new information I consists entirely of subset and aggregate constraints, we refer to $I = SAA$ or $I' = S'AA'$ as fully decomposable

constraints.

B. Decomposition method of Computing $p \circ S'$

From (20) we know that the posterior conditionals of $r = p \circ S'$ can be obtained by solving the minimum cross-entropy problems $(p^* \mathcal{D}_j) \circ S_j$ on each subset. If $\Psi r = \Psi p$ were true, then one could construct the full posterior $r = p \circ S'$ without solving a minimum cross-entropy problem on the full space \mathcal{D} . Unfortunately, $\Psi r = \Psi p$ does not hold in general -- for an example, see the discussion following Property 8 in [13]. Nevertheless, we shall show in this Section that it is simple to compute the posterior aggregate distribution Ψr from the subset results $(p^* \mathcal{D}_j) \circ S_j$ and that one can therefore construct $p \circ S'$ without solving a minimum cross-entropy problem on the full space \mathcal{D} .

From (11) and (18) it follows that $r = p \circ S'$ is given by

$$r(\underline{x}) = p(\underline{x}) \exp \left(-\theta - \sum_{j=1}^M \sum_{i=1}^{K_j} \alpha_{ij} f_{ij}(\underline{x}) \right), \quad (27)$$

where θ and α_{ij} are Lagrangian multipliers. Using (9) and (19) it follows that the conditional density in the subset \mathcal{D}_j is

$$\begin{aligned} r^* \mathcal{D}_j &= r_j^{-1} r(\underline{x} \in \mathcal{D}_j) \\ &= r_j^{-1} p(\underline{x}) \exp \left(-\theta - \sum_{i=1}^{K_j} \alpha_{ij} s_{ij}(\underline{x}) \right) \end{aligned} \quad (28)$$

where $r_j = [\Psi r]_j$. Now the conditional prior is $p^* \mathcal{D}_j = p(\underline{x})/p_j$, where $p_j = [\Psi p]_j$, and it follows from (11) and (17) that

$$(p^* \mathcal{D}_j) \circ S_j = p_j^{-1} p(\underline{x}) \exp \left(-\xi_j - \sum_{i=1}^{K_j} \eta_{ij} s_{ij}(\underline{x}) \right). \quad (29)$$

Owing to weak subset independence (20), it follows from (28)-(29) that

$$\alpha_{ij} = \eta_{ij} \text{ and}$$

$$r_j = p_j \exp(-\theta + \xi_j), \quad (30)$$

provided that the constraint functions s_{ij} in (17) are linearly independent.

Now, on the right side of (30), the p_j are known from the prior and the ξ_j are the normalization Lagrangian multipliers from the subset conditional

densities $(p^* \mathcal{D}_j) \circ S_j$. Since $e^{-\theta}$ is just a normalization factor

($\sum_j r_j = 1$), it follows that one can compute the posterior aggregate probabilities r_j from (30) using p and the results of $(p^* \mathcal{D}_j) \circ S_j$. From the posterior aggregates and the posterior conditionals, one can then construct the full posterior $r = p \circ S'$ since $r(x \in \mathcal{D}_j) = r_j (r^* \mathcal{D}_j)$. This result can also be seen by noting that the multipliers $\alpha_{ij} = \eta_{ij}$ and θ in (27) are known from (29)-(30) and the normalization requirement $\sum_j r_j = 1$.

C. Cross-Entropy Minimization Given Fully Decomposable Constraints

In this subsection, we prove the following new properties that hold in the case of fully decomposable subset and aggregate constraints:

$$p \circ (S' \wedge A') = (p \circ S') \circ A' \quad (31)$$

$$(p \circ (S' \wedge A'))^* \mathcal{D}_j = (p^* \mathcal{D}_j) \circ S_j \quad (32)$$

$$\Psi(p \circ (S' \wedge A')) = (\Psi(p \circ S')) \circ A \quad (33)$$

For convenience, we define

$$r = p \circ S' \quad (34a)$$

$$u = (p \circ S') \circ A' \quad (34b)$$

$$q = p \circ (S' \wedge A') \quad (34c)$$

In these terms, the following relations also hold:

$$H[q, p] = H[r, p] + H[\Psi q, \Psi r] \quad (35)$$

$$H[q, p] = H[\Psi q, \Psi p] + \sum_{j=1}^M [\Psi q]_j H[q^* \mathcal{D}_j, p^* \mathcal{D}_j] \quad (36)$$

Discussion: For arbitrary constraints I_a and I_b , $p(I_a \wedge I_b)$ is not in general the same as $(p \circ I_a) \circ I_b$. One way to see this is to note that, while $p \circ I_a$ is guaranteed to satisfy I_a , $(p \circ I_a) \circ I_b$ is guaranteed only to satisfy I_b — I_a may remain satisfied but only in special cases. Fully decomposable constraints are an example of such a case. For another example, see Property 4 in [13]. Eq. (32) shows that the conditional densities $q^* \mathcal{D}_j$ can be obtained by solving the subset problems $(p^* \mathcal{D}_j) \circ S_j$. In the previous subsection we showed how one can use the resulting solutions to find the aggregate distribution $\Psi r = \Psi(p \circ S')$. Since (33) shows that one can find the aggregate distribution Ψq by solving a minimum cross-entropy problem using Ψr as a prior, it follows that one can construct q without solving a minimum cross-entropy problem on the full space \mathcal{D} .

Proofs: We begin with explicit expressions for the densities (34). The density r is given by (27). Hence, the density u is given by

$$u(\underline{x}) = p(\underline{x}) \exp \left(-\theta - \sum_{j=1}^M \sum_{i=1}^{K_j} \alpha_{ij} f_{ij}(\underline{x}) - \phi - \sum_{i=1}^{K_a} \mu_i a_i(\underline{x}) \right), \quad (37)$$

where ϕ is a Lagrangian multiplier corresponding to a normalization constraint and where μ_i are multipliers corresponding to the constraints (23). The density q is given by

$$q(\underline{x}) = p(\underline{x}) \exp \left(-\lambda - \sum_{j=1}^M \sum_{i=1}^{K_j} \beta_{ij} f_{ij}(\underline{x}) - \sum_{i=1}^{K_a} \delta_i a_i(\underline{x}) \right), \quad (38)$$

where λ , β_{ij} , and δ_i are Lagrangian multipliers corresponding respectively to (4), (19), and (23).

Now it follows from (9), (19), (24), and (38) that the conditional density $q^* \mathcal{D}_j$ is

$$\begin{aligned}
[q * \mathcal{D}_j](x) &= q_j^{-1} q(x \in \mathcal{D}_j) \\
&= q_j^{-1} p(x) \exp \left(-\lambda - \sum_{i=1}^{K_j} \beta_{ij} s_{ij}(x) - \sum_{i=1}^{K_a} \delta_i a_{ij} \right), \quad (39)
\end{aligned}$$

where $q_j = [\psi q]_j$. Similarly, it follows from (28) that

$$[r * \mathcal{D}_j](x) = r_j^{-1} p(x) \exp \left(-\theta - \sum_{i=1}^{K_j} \alpha_{ij} s_{ij}(x) \right). \quad (40)$$

Now (39) has the form

$$[q * \mathcal{D}_j](x) = A_j p(x) \exp \left(- \sum_{i=1}^{K_j} \beta_{ij} s_{ij}(x) \right) \quad (41)$$

with

$$A_j = q_j^{-1} \exp \left(-\lambda - \sum_{i=1}^{K_a} \delta_i a_{ij} \right), \quad (42)$$

and (40) has the form

$$[r * \mathcal{D}_j](x) = B_j p(x) \exp \left(- \sum_{i=1}^{K_j} \alpha_{ij} s_{ij}(x) \right) \quad (43)$$

with

$$B_j = r_j^{-1} \exp(-\theta) \quad (44)$$

Since both (41) and (43) have the same form, satisfy (17), and integrate to unity, it follows that they are equal everywhere on \mathcal{D}_j . Thus

$$(p \circ (S' \wedge A')) * \mathcal{D}_j = (p \circ S') * \mathcal{D}_j \quad (45)$$

holds, as well as $\alpha_{ij} = \beta_{ij}$ and $A_j = B_j$. Eq. (32) then follows

directly from (20) and (45).

Since $A_j = B_j$ holds, (42) and (44) yield

$$q_j = r_j \exp \left(\theta - \lambda - \sum_{i=1}^{K_a} \delta_i a_{ij} \right) . \quad (46)$$

Now, we have

$$[(\psi r) \circ A]_j = r_j \exp \left(-\gamma - \sum_{i=1}^{K_a} \epsilon_{ja} a_{ij} \right) , \quad (47)$$

where γ and ϵ_j are Lagrangian multipliers corresponding respectively to a normalization constraint and to (22). Eq. (46) also satisfies these constraints. Since it has the same form as (47), it follows that the two equations are equal; that is, $\psi q = (\psi r) \circ A$ holds, which is (33).

Now $(\psi(p \circ S')) \circ A = \psi((p \circ S') \circ A')$ holds as a consequence of subset aggregation (25) — just substitute $p \circ S'$ for p in (25). It follows from (33) that

$$\psi(p \circ (S' \wedge A')) = \psi((p \circ S') \circ A') \quad (48)$$

holds. If it is also true that

$$(p \circ (S' \wedge A')) * \mathcal{D}_j = ((p \circ S') \circ A') * \mathcal{D}_j \quad (49)$$

or $q * \mathcal{D}_j = u * \mathcal{D}_j$ holds, then (31) follows immediately. To see that (49) does indeed hold, we use (9), (19), (24), and (37) to express $u * \mathcal{D}_j$ as

$$[u * \mathcal{D}_j](x) = u_j^{-1} p(x) \exp -\theta - \varphi - \sum_{i=1}^{K_j} \alpha_{ij} s_{ij}(x) - \sum_{i=1}^{K_a} \mu_i a_{ij} ,$$

where $u_j = [\psi u]_j$. This has the form

$$[u * \mathcal{D}_j](x) = c_j p(x) \exp \left(- \sum_{i=1}^{K_j} \alpha_{ij} s_{ij}(x) \right) \quad (50)$$

with

$$C_j = u_j^{-1} \exp \left(-\theta - \varphi - \sum_{i=1}^{K_a} \lambda_i a_{ij} \right) .$$

Now (43) and (50) differ only in the leading factors B_j or C_j . Since they both integrate to unity, it follows that $B_j = C_j$ and $r^* \underline{D}_j = u^* \underline{D}_j$ or $(p \circ S')^* \underline{D}_j = ((p \circ S') \circ A')^* \underline{D}_j$. Eq. (49) then follows from (45). This completes the proofs of (31)-(33).

To prove (35)-(36), we note that, since the right hand sides of (17), (18), (22), and (23) are zero, it follows from (12) that

$$\xi_j = -H[(p^* \underline{D}_j) \circ S_j, p^* \underline{D}_j] = -H[q^* \underline{D}_j, p^* \underline{D}_j] \quad (51)$$

$$\theta = -H[r, p] \quad (52)$$

$$\lambda = -H[q, p] \quad (53)$$

all hold, where ξ_j , θ , and λ are Lagrangian multipliers from (29), (27), and (38). Eq. (46) yields

$$\begin{aligned} H[\underline{\psi}_q, \underline{\psi}_r] &= \sum_{j=1}^M q_j \log(q_j / r_j) \\ &= \theta - \lambda - \sum_{i=1}^{K_a} \delta_i \sum_{j=1}^M q_j a_{ij} \\ &= \theta - \lambda \end{aligned}$$

Eq. (35) then follows from (52)-(53). Since (21) holds in general, (36) follows directly by substitution of q for r . Alternatively, we use (32) to equate the right side of (29) with the right side of (39). Since $\alpha_{ij} = \eta_{ij} = \beta_{ij}$ holds -- as pointed out in the discussion following (29) and (45) -- it follows that

$$\log(q_j/p_j) = \xi_j - \lambda - \sum_{i=1}^{K_a} \delta_i a_{ij}$$

holds. After multiplying this by q_j and summing, we obtain (36) by substituting (51) and (53).

IV. COMPUTATIONAL RESULTS

In this Section we present some numerical examples to demonstrate the savings that can be obtained using the decomposition method of computing $q = p \circ (S' \wedge A')$. We compare the following two methods of computing q :

Method A (decomposition).

- a) obtain the posterior conditional densities $q^* \mathcal{D}_j$ by computing $(p^* \mathcal{D}_j) \circ S_j$ -- See (32);
- b) compute the aggregate distribution $\Psi(p \circ S')$ using the results of $(p^* \mathcal{D}_j) \circ S_j = (p \circ S')^* \mathcal{D}_j$ as explained in Section III(B);
- c) obtain the posterior aggregate distribution Ψq by computing $(\Psi(p \circ S')) \circ A$ -- See (33); and
- d) combine $q^* \mathcal{D}_j$ and Ψq to obtain the full posterior q .

Method B (full space).

- a) obtain q by solving $p \circ (S' \wedge A')$ directly.

In order to compare the two methods, we computed examples for several values of N , M , K , and K_a , where

- N = total size of discrete state space D ,
- M = number of subsets, each with size N/M ,

K = number of constraints per subset, and

K_a = number of aggregate constraints.

In these terms, the decomposition method requires solving M minimum cross-entropy problems of dimension N/M each with K constraints followed by one problem of dimension M with K_a constraints. The full space method requires solving a single minimum cross-entropy problem of dimension N with $MK + K_a$ constraints. Since the computational complexity of minimum cross-entropy problems grows rapidly with the dimension of the problem, the decomposition method can lead to considerable savings.

Specifically, we used the following procedure, where $U[0,1]$ refers to a psuedo-random number uniformly distributed on $[0,1]$:

1. Construct a random "unknown" distribution q^\dagger by picking N $U[0,1]$ values and normalizing.
- 2) For each of the subsets \mathcal{D}_j , $j = 1, \dots, M$, construct K random subset constraints S_j by picking N/M $U[0,1]$ values as coefficients for each constraint and computing the expectations of $q^\dagger * \mathcal{D}_j$.
- 3) Construct K_a random aggregate constraints A by picking M $U[0,1]$ values as coefficients for each constraint and computing the expectations of Ψq^\dagger .
- 4) Construct a random prior p by picking N $U[0,1]$ values and normalizing.
- 5) Compute the posterior $q = p \circ (S'AA')$ by both methods and compare execution times. For the computations, we used a slightly modified version of the APL function MINCROSSENT described in [14]. The modification made the normalization constant $\exp(-\lambda)$ in (11) available after the function call as a global variable.

We fixed the subset size at $N/M = 10$ and the number of subset constraints per subset at $K = 5$. We computed results for $M = 2, 4, 6, 8$ subsets with

$K_a = M/2$ aggregate constraints in each case. For each set of values N , M , K , and K_a , we repeated the foregoing procedure four times and averaged the resulting execution times, which are expressed as seconds of execution time on an IBM 370/158 processor. The results are summarized in Table I. For the case of 80 states and 44 total constraints, the decomposition method was more than ten times faster.

Table I. Comparison of Decomposition and Full-Space Methods

No. of Subsets	Total No. of States	Total No. of Constraints	Method A (Decomp.) (secs.)	Method B (Full) (secs.)	Ratio of B to A
2	20	11	.45	.57	1.3
4	40	22	.86	2.9	3.4
6	60	33	1.3	8.9	6.8
8	80	44	1.9	22	11.6

V. ACKNOWLEDGEMENTS

The author thanks J. Abrahams, R. Johnson, and H. Vantilborgh for helpful discussions.

REFERENCES

1. S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
2. E.T. Jaynes, "Information Theory and Statistical Mechanics I", Phys. Rev. 106, 1957, pp.620-630.
3. W.M. Elsasser, "On Quantum Measurements and the Role of the Uncertainty Relations in Statistical Mechanics," Phys. Rev. 52, (Nov. 1937), pp. 987-999.
4. J.E. Shore and R.W. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," IEEE Trans. Inf. Theory, vol. IT-26, no. 1, Jan. 1980.
5. J.E. Shore, "Minimum Cross-Entropy Spectral Analysis," IEEE Trans. Acoustics, Speech, & Signal Proc., to appear.
6. R.M. Gray, A.H. Gray, Jr., G. Rebolledo, and J.E. Shore, "Rate-Distortion Speech Coding With a Minimum Discrimination Information Distortion Measure," submitted to IEEE Trans. Information Theory.
7. J.E. Shore and R.M. Gray, "Minimum Cross-Entropy Pattern Classification and Cluster Analysis," NRL Memorandum Report 4207, Naval Research Laboratory, Washington, D.C. 20375, April 1980, submitted to IEEE Trans. Pattern Anal. & Mach. Intell.
8. J.E. Shore, "Derivation of Equilibrium and Time-Dependent Solutions to M/M/oo//N and M/M/oo Queueing Systems Using Entropy Maximization," Proceedings 1978 National Computer Conference, AFIPS 47, pp. 483-487.
9. J.E. Shore, "Information Theoretic Approximations for M/G/1 Queueing Systems," in D.G. Lainiotis and N.S. Tzannes (eds.), Advances in Communications, Dordrecht, Holland, D. Reidel Publishing Co., 1980, pp. 321-332.
10. Y. Bard, "A Model of Shared DASD and Multipathing," Technical Report G320-2130, IBM Scientific Center, Cambridge, Mass., May 1980.
11. Y. Bard, "Estimation of State Probabilities Using the Maximum Entropy Principle," IBM J. Res. Develop. 24, 190, pp. 563-569.
12. I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," Ann. Prob., vol. 3, pp. 146-158, 1975.
13. J.E. Shore and R.W. Johnson, "Properties of Cross-Entropy Minimization," IEEE Trans. Information Theory, to appear (July, 1981).

14. R.W. Johnson, "Determining Probability Distributions by Maximum Entropy and Minimum Cross-Entropy," Proceedings APL79, pp. 24-29.
15. D.V. Gokhale and S. Kullback, The Information in Contingency Tables, Marcel Dekker, New York, 1978.
16. N. Agmon, Y. Alhassid, and R.D. Levine, "An Algorithm for Finding the Distribution of Maximal Entropy," J. Comput. Phys. 30, 1979, pp. 250-258.
17. R.W. Johnson, "Axiomatic Characterization of the Directed Divergences and Their Linear Combinations," IEEE Trans. Inf. Theory, vol. IT-25, no. 6, pp. 709-716, Nov. 1979.

